Word Sense Disambiguation of Text by Association Methods: A Comparative Study

Richard F.E. Sutcliffe
Bronwyn E.A. Slater

Department of Computer Science
and Information Systems
University of Limerick
Limerick, Ireland

+353 61 333644 Ext 5006
+353 61 330876 (FAX)

sutcliffer@ul.ie
slaterb@ul.ie

Word sense disambiguation is vital to accurate text analysis. We have replicated two well known methods du Lesk (1986) and Ide and Veronis (1990), and have conducted experiments using both methods on a corpus of sentences. We also carried out experiments to test our theory that syntactic tagging would improve results. V have made some interesting discoveries. Firstly, syntactic tagging improves the performance of the disambig algorithms. Secondly, the Ide and Veronis method performs only superficially better than the Lesk method. Thirdly, the performance of a particular algorithm is heavily dependent on the way in which it is measured.

Palabras: Diccionarios en formato maquina, ambiguedad de vocablos, cadenas neuronas artificial, la propaga de activaridad, medicion de funcionamiento en linguistica computacional.

Topic Areas: computational lexicology, word sense disambiguation, neural networks, objective performance metrics in NLP systems.

1 Introduction

The ability to disambiguate the senses of each word in a text is vital if its meaning is to be determine In '*Animal Farm*' by George Orwell for example, 87% of words have more than one sense. With the a machine readable dictionaries (MRDs) various methods have been proposed to disambiguate words a Two well-known techniques are the Lesk method (Lesk, 1986) and the Cottrell-Veronis-Ide (CVI) m (Veronis and Ide, 1990). Both of these exploit the idea that the correct senses of a pair of words in a will be semantically related and that this can be detected using their definitions. We have replicated using the same dictionary and test corpus in order to determine their relative performance in an iden addition we have investigated whether syntactic category tagging can be used to improve performanc eliminating word senses which are of the wrong category.

2 Contrasts with related work

We are aware of the work of Guthrie et al. (1990) in which disambiguation was carried out by assoc semantic codes of the Longmans Dictionary of Contemporary English (LDOCE), yielding approxim success. In addition Schuetze (1993) and Church and Yarowsky (1992) have both developed corpus which yield very high performance rates. In these the target word in its context is compared with va previously analysed occurrences of the word which have been grouped by sense, using statistical me both approaches are very interesting, they may not be applicable to a given disambiguation task. Tl approach is specific to the LDOCE because it depends on the topic codes which are associated with sense in that dictionary. On the other hand the Schuetze and Gale method can not readily be linkec dictionary word senses which makes it hard to use in conjunction with, say, a semantic lexicon deri dictionary. It is for these reasons that we wished to investigate the performance of the Cottrell-Ver Lesk methods.

ɜ Disambiguation Algorithms

ıe Disambiguation Task

ɔrd sense disambiguation task our objective is to assign to each word in a text an appropriate sense chosen particular MRD. Thus to disambiguate "pen paper" relative to the Merriam-Webster Compact Electronic nary we choose sense three of 'pen' ("tool for writing with ink"), and sense one of 'paper' ("pliable substance ɔ write or print on, to wrap things in, or to cover walls").

he Lesk Method

esk disambiguation method involves the use of frequency counts in computing the preferred sense of each in the input phrase or sentence. Firstly, all the sense definitions of each word in the input are looked up in ıctionary. Analysis then proceeds by discarding each word in a sense definition which does not occur in any definition. Each remaining word in a definition is converted to its root inflection. A count is made of the ɔer of times it occurs in other definitions and that count is then associated with the word wherever it occurs. A is then determined for each sense definition by computing the product of the word scores within it. Finally, word is disambiguated by choosing the sense which has the highest score.

The Cottrell-Ide-Veronis Method

ɹell-Ide-Veronis disambiguation is similar in spirit to the Lesk method but uses a spreading activation network ı two-way arcs. There are two types of node in the network, word nodes and sense nodes. The network is ıted by first allocating one word node to each content word in the input (function words are eliminated). Thus "pen paper" one word node is allocated for 'pen' and another for 'paper'. Each word node is connected by itatory arcs to sense nodes, one for each semantic sense of the word as defined in the dictionary. Thus we might

nodes for a word is strongly interconnected by inhibitory arcs to form a winner-take-all network. Eac
is then connected to one word node for each word occurring in that sense definition, converted to its
Thus pen3 would be connected to 'tool', 'write' and 'ink'. These links are excitatory. There is only one
each word node in the network. Thus if a word occurs in more than one definition, several sense noc
connected to it. Because these nodes join different parts of the network, the system can capture the s
between senses of different words in the input.

The activation functions used in the network are very standard. The activation at time $t + 1$, $a_i(t + 1)$
follows:

$$a_i(t + 1) = a_i(t) + s_i - \delta$$

The squashed net input s_i is defined by

$$s_i = n_i(1 - a_i) \quad \text{when } n_i > 0$$

$$s_i = n_i a_i \quad \text{when } n_i < 0$$

where the net input to node i, n_i is

$$n_i = \Sigma w_{ji} a_j$$

Decay δ is given by

$$\delta = D_1(a_i - D_2)$$

where D_1 and D_2 are constants.

After the network has been created, the cycling phase begins. The activation of the input word nodes is set to 0.2 and the network is run until a situation of stability has been reached. Words which occur in sense definitions of several different input words will tend to become more active because they receive input from more than one part of the network. As a result they will tend to reinforce the sense nodes to which they are connected, thus pushing down competing senses. Disambiguation is accomplished by choosing from each winner-take-all network the sense which has the highest activation.

The network described above is of height one (CVI-1). A CVI-2 network can be created by taking each word node which occurs at the bottom of the network and creating further nodes for it. Firstly, we create a sense node for each sense of that word in the dictionary. Secondly, we add word nodes under each sense node corresponding to the words which occur in the definition of that sense, just as before. In general, a CVI network of any height can be created by repeating this process.

4 Disambiguation Experiments

4.1 Corpus Creation and Sense Tagging

We created a corpus of 100 sentences from '*Animal Farm*' by George Orwell. Function words were eliminated. Each word was then disambiguated manually by two human subjects. During the disambiguation session, a subject was presented with a complete sentence on the screen together with the appropriate definitions from the Merriam-Webster Compact Electronic Dictionary. They then selected zero, one or more senses for each word which they considered appropriate for its use in that context. The results of each session were saved in a file. The 'correct' set of senses for each word in a given sentence was then created by taking the intersection of the sets created for it by the pair of subjects. Two disambiguation experiments were then carried out on the corpus.

4.2 Experiment 1 : Corpus not syntactically tagged

In the first experiment the Lesk, CVI-1, CVI-2 and CVI-3 methods were used in turn to disambiguate the
The results produced by each algorithm were compared with those indicated by the human subjects. Thes
summarised in Table One.

4.3 Experiment 2 : Corpus syntactically tagged

In the second experiment we tagged the corpus for syntactic category using the Brill Tagger (Brill, 1993)
assigns a syntactic category to each word in the text with high accuracy. We then ran the Lesk, CVI-1, C
CVI-3 models again, this time utilising the syntactic information to restrict possible word sense choices.
accomplished by only considering word senses produced by the algorithms which were of the same categ
predicted by the tagger. Thus for example if the word to be disambiguated is 'dog' as a noun, we do not
algorithm to select a verb sense of dog. In the Lesk method we choose the word sense of the correct cate
has the highest score. This is not necessarily the sense which has the highest score overall, which could
'dog' as a verb. Similarly in a CVI method we select the sense node which has the highest activation an
right category. The results of the experiment are summarised in Table Two.

4.4 Results

Before discussing the results of our study we define the terms used in the tables:

Content words are defined to be those of category noun, verb, adjective or adverb. All other words are
to be function words and are thus excluded.

An **ambiguous word** has more than one sense in the dictionary while an **unambiguous word** has only
The row marked **Total correct** is a count of the ambiguous words which were disambiguated correctly
of all the unambiguous words.

Table One : Experiment 1 - not syntactically tagged

	Lesk	CVI-1	CVI-2	CVI-3
Total sentences	100	100	100	26
Total words	2647	2647	2647	679
Total content words	1094	1094	1094	289
Total ambiguous words	954	954	954	250
	(87%)	(87%)	(87%)	(86%)
Total unambiguous words	140	140	140	39
	(13%)	(13%)	(13%)	(14%)
Total correct	422	660	651	168
(includes unambiguous words)	(39%)	(60%)	(59.5%)	(58%)
Total ambiguous correct	282	520	511	129
	(30%)	(55%)	(54%)	(52%)
Total ambiguous isolated	413	413	0	0
	(43%)	(43%)		
Total ambiguous non-isolated	537	537	954	250
	(57%)	(57%)	(100%)	(100%)
Total ambiguous	281	286	511	129
non-isolated correct	(52%)	(53%)	(54%)	(52%)

Table Two : Experiment 2 - syntactically tagged

	Lesk	CVI-1	CVI-2
Total sentences	100	100	100
Total words	2647	2647	2647
Total content words	1013	1013	1013
Total ambiguous words	878	878	878
	(87%)	(87%)	(87%)
Total unambiguous words	135	135	135
	(13%)	(13%)	(13%)
Total correct	406	732	728
(includes unambiguous words)	(40%)	(72%)	(72%)
Total ambiguous correct	271	597	593
	(31%)	(68%)	(68%)
Total ambiguous isolated	463	463	0
	(53%)	(53%)	
Total ambiguous non-isolated	415	415	878
	(47%)	(47%)	(100%)
Total ambiguous	271	269	593
non-isolated correct	(65%)	(65%)	(68%)

Total ambiguous correct is the number of ambiguous words which were disambiguated correctly.

Isolated words are those whose definitions share no words with other definitions in the sentence being disambiguated. By definition, such words can not possibly be disambiguated by either an Ide or a Lesk u

The row marked **Total ambiguous isolated** is a count of the ambiguous words which are isolated.

Total ambiguous non-isolated is thus a count of the ambiguous words which are in principle disambigu the methods.

Finally, **total ambiguous non-isolated correct** is a count of the disambiguatable words which were corre chosen by the methods. This is the true measure of performance of the algorithms.

The main findings can be summarised as follows:

Firstly, we note from the 'total ambiguous non-isolated correct' figures that both the Lesk and CVI-1 alg perform equally well in both experiments. This figure represents those words which can in fact be disan We should note that the Lesk method is incapable of choosing a sense of a word whose definition is isol However, the CVI method will always choose some sense of a word, even if its definition is isolated. Th for the apparent superiority of the CVI method over the Lesk method (see 'total ambiguous correct' figu experiments. In experiment 1 the 'total ambiguous correct' figure for Lesk was 30% compared with 55% 1. The 25% difference here can be attributed to the CVI-1 method choosing senses by chance. In expe shortfall is 37%, again due to the CVI-1 method choosing senses by chance.

Secondly, syntactic tagging did increase the performance of all algorithms, in all categories of measure

Thirdly, CVI-3 networks never performed better than CVI-2 networks. In addition CVI-3 networks we considerably larger in size, comprising around 3000-4000 nodes and 10,000 bidirectional arcs. This su the 'interesting' words occur in the more immediate dictionary definitions rather than in those at a deep Thus CVI-3 networks may not be worth the extra space and time requirements which they incur.

4.5 Conclusions

Our conclusions may be summarised as follows:

- Both the Lesk and CVI methods perform comparably well. However, in the case of words which cannot be disambiguated by the association methods the CVI algorithm has the capability of choosing senses at random. Th Lesk method does not have this capability.

- We would tend to favour the CVI-1 method as it gives the best performance while the CVI-2 and CVI-3 methods yield no improvements and are much slower. From this we conclude that the 'interesting' words are occurring at the top levels of the disambiguation network.

- Syntactic tagging improves the chances of a correct sense being chosen, for both the Lesk and CVI algorithms.

- Only the top rate of performance achieved here (72%) is comparable to that reported in the word-pair study conducted by Ide and Veronis (1990).

- The way in which performance is measured makes a large difference to the results. In particular, including 1-way ambiguous words artificially boosts results. Eliminating the senses which were chosen by chance is also important. These occur where words are isolated and are thus intrinsically undisambiguatable.

Factors not investigated in this study include:

- **The number of words disambiguated at a time.** The use of whole sentences makes the disambiguation task more difficult but it seems a likely way in which an algorithm would be used in a text processing application.

- **The domain of the corpus.** The particular application domain in which the disambiguation is to be used may well affect results. In addition, higher levels of performance can undoubtedly be obtained in restricted contexts where domain-specific word-sense frequency data can be exploited. For example in a computer manual 'file' almost certainly means a computer file.

- **Criteria for the selection of test sentences.** Undoubtedly the size of the corpus and its composition in terms sentence length, proportion of function words and so on will affect results.

- **The effect of the dictionary used.** We used the same dictionary for all trials, namely the CED. It is possible however that other dictionaries could give a higher level of performance overall or that they particularly suit a given algorithm.

Clearly there are many interesting avenues for this work and we are currently engaged in researching some of them.

5 References

Brill, E. (1993). *A Corpus-Based Approach to Language Learning*. Doctoral Dissertation, Department of Computer and Information Science, University of Pennsylvania.

Cottrell, G.W. (1985). *A Connectionist Approach to Word Sense Disambiguation*. Doctoral Dissertation, Department of Computer Science, University of Rochester.

Gale, W. A., Church, K. W., Yarowsky, D. (1993). A Method for Disambiguating Word Senses in a Large Corpus. *Computers and the Humanities*, **26**(5-6), 415-439.

Guthrie, L., Slator, B., Wilks, Y., Bruce, R. (1990). Is there content in Empty Heads? *Proceedings of the 13th International Conference on Computational Linguistics (COLING-90), Helsinki, Finland, 3,pp.138-143.*

Ide, N. M, & Veronis, J. (1990). Very Large Neural Networks for Word Sense Disambiguation. *Proceedings of the European Conference on Artificial Intelligence, ECAI'90, Stockholm, August 1990.*

Lesk, M. (1986). Automated Word Sense Disambiguation using Machine-Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone. *Proceedings of the 1986 SIGDOC Conference.*

Schuetze, H. (1993). Word Space. In S. J. Hanson, J. D. Cowan and C. L. Giles (Eds.), *Advances in Neural Information Processing Systems 5.* San Mateo CA: Morgan Kaufmann.